Snow Model Verification Using Ensemble Prediction and Operational Benchmarks

KRISTIE J. FRANZ*

Department of Civil and Environmental Engineering, University of California, Irvine, Irvine, California

TERRI S. HOGUE

Department of Civil and Environmental Engineering, University of California, Los Angeles, Los Angeles, California

SOROOSH SOROOSHIAN

Department of Civil and Environmental Engineering, University of California, Irvine, Irvine, California

(Manuscript received 14 November 2007, in final form 27 May 2008)

ABSTRACT

Hydrologic model evaluations have traditionally focused on measuring how closely the model can simulate various characteristics of historical observations. Although advancing hydrologic forecasting is an often-stated goal of numerous modeling studies, testing in a forecasting mode is seldom undertaken, limiting information derived from these analyses. One can overcome this limitation through generation, and subsequent analysis, of ensemble hindcasts. In this study, long-range ensemble hindcasts are generated for the available period of record for a basin in southwestern Idaho for the purpose of evaluating the Snow–Atmosphere–Soil Transfer (SAST) model against the current operational benchmark, the National Weather Service's (NWS) snow accumulation and ablation model SNOW17. Both snow models were coupled with the NWS operational rainfall runoff model and ensembles of seasonal discharge and weekly snow water equivalent (SWE) were evaluated. Ensemble predictions from both the SAST and SNOW17 models were better than climatology forecasts, for the period studied. In most cases, the accuracy of the SAST-generated predictions was similar to the SNOW17-generated predictions, except during periods of significant melting. Differences in model performance are partially attributed to initial condition errors. After updating the SWE state in the snow models with the observed SWE, the forecasts were improved during the first 2–4 weeks of the forecast window and the skills were essentially equal in both forecasting systems for the study watershed. Climate dominated the forecast uncertainty in the latter part of the forecast window while initial conditions controlled the forecast skill in the first 3–4 weeks of the forecast. The use of hindcasting in the snow model analysis revealed that, given the dominance of the initial conditions on forecast skill, streamflow predictions will be most improved through the use of state updating.

1. Introduction

Hydrologic analyses seldom address a model's forecasting ability, despite this being an often-stated motivation in many hydrologic modeling studies. Recently, Welles et al. (2007) identified forecast verification as an obvious gap in hydrologic research aimed at improving hydrologic forecasting. The hydrologic research community has generally focused on the validation of new techniques through simulation, while forecast verification has yet to be widely used. Forecast verification provides an objective means to guide future research aimed at improving hydrologic forecasts (Welles et al. 2007). Perkins (1988), Rango and Martinec (1994), and Gomez-Landesa and Rango (2002) included forecasting evaluation in their snow model analyses. Several of these studies were conducted in real time, either by, or in cooperation with, operational agency personnel.

Researchers are limited as they may not be able to evaluate their techniques over the range of conditions that will be found in the forecast setting. However,

* Current affiliation: Department of Geological and Atmospheric Sciences, Iowa State University, Ames, Iowa.

Corresponding author address: Dr. Kristie Franz, Dept. of Geological and Atmospheric Sciences, Iowa State University, 3023 Agronomy Hall, Ames, IA 50011.
E-mail: kfranz@iastate.edu

DOI: 10.1175/2008JHM995.1

without analysis of the predictive skills of the model during research and development, there may be insufficient information to rigorously assess the possible outcome of applying new techniques operationally. Although all contingencies of an operational forecast environment cannot be addressed, hindcasting provides valuable insight on alternative methods and can easily be applied within the research setting. Hindcasting has been used to evaluate the application of climate signals for ensemble weighting (Werner et al. 2004; Hamlet and Lettenmaier 1999), to examine the use of climate forecast model output as forcing for hydrologic prediction models (Carpenter and Georgakakos 2001; Wood et al. 2005), and to create sufficient forecast samples to investigate forecasting systems and verification methods (Wood et al. 2002; Franz et al. 2003; Bradley et al. 2004). Recent publications have also used hindcasting to investigate the use of numerical weather predictions in hydrologic forecasting (Westrick et al. 2002; Roulin and Vannitsem 2005; Werner et al. 2005; Clark and Hay 2004).

The work presented here evaluates the use of ensemble streamflow prediction (ESP) hindcasts as the definitive step in the investigation of the Snow–Atmosphere–Soil Transfer model (SAST; Jin et al. 1999a), a snow energy balance model, for use in the National Weather Service River Forecasting System (NWSRFS). Kirchner et al. (1996) states that evaluation of a model requires three critical elements: a performance criterion, a benchmark, and an outcome. Performance criterion refers to the ability to match the desired variable being modeled; in this instance the variables of interest are simulated snow water equivalent (SWE), melt, and discharge. The benchmark is an alternative to the model being evaluated. Given forecasting as the proposed application of the SAST, our benchmark is identified as the operational National Weather Service (NWS) SNOW17 model (Anderson 1973), a temperature-based snow model. The outcome describes how the model performs with respect to the benchmark; the performance of interest is producing ensemble streamflow predictions (ESPs) as reliable as, or better than, the benchmark. Ensemble streamflow predictions are expected to be a key component of the planned NWS Advanced Hydrologic Predictions Services (AHPS; McEnery et al. 2005); thus, a new forecast snow model will have to be viable within the scope of this forecasting method.

Implementation of a forecast model is complicated, primarily due to the lack of dialogue between the researcher and operational forecaster to develop and frame algorithms that will work in the "operational world" (Rango 1989). The NWS is making efforts to improve the communication between forecasters and researchers through new projects such as the Hydrologic Ensemble Predictions EXperiment (HEPEX; Franz et al. 2005) and development of the Advanced Hydrologic Prediction Services (AHPS; McEnery et al. 2005). Infusion of new science and technology will be limited by the inability to test alternative methods with other components of the forecast system. Recently, the Community Hydrologic Prediction System (CPHS) has been proposed (Schaake et al. 2006) which will have an open software infrastructure allowing new components to be "plugged in" to the system more easily. However, until CHPS is developed and well tested, the NWSRFS will continue to be the official hydrologic forecast system in the United States. Any research advances will have to be made compatible with the current NWSRFS in order to be transferred to operations in the near future.

In a prior study, the SAST was coupled to the NWS Sacramento Soil Moisture Accounting Model (SACSMA; Burnash et al. 1973) and compared to the existing NWS SNOW17–SACSMA modeling system based on seasonal snowpack and discharge simulations (Franz et al. 2008). The study area included two nested basins in the Reynolds Creek Experimental Watershed (RCEW), Idaho (Slaughter et al. 2001). Both snow models were shown to be equally skillful for most years, with the SNOW17 being less prone to large overestimations of seasonal SWE and less sensitive to input data uncertainty. On average, the SAST more accurately matched the timing of completion of the snowpack melt, but tended to overestimate SWE and rapidly melt snow in the spring leading to overestimated peak discharge during several years (when the SAST is coupled to the SACSMA). The results of the study suggested the potential for the implementation of the SAST in seasonal streamflow prediction within the NWSRFS given further understanding of model, parameter, and data uncertainty.

In the current work, we provide a framework for a more rigorous evaluation of the alternative snow model performances by incorporating ensemble prediction and operational benchmarks to evaluate the skill for coupled snow–runoff forecasting. Thirteen years of simulated historical ESP outlooks (hindcasts) of SWE and streamflow are created using both the SAST and SNOW17. Hindcasts are analyzed using common forecast verification statistics. The ensemble prediction method of the NWS provides a quantitative assessment of the impact of future climate uncertainties on predicted streamflow. The SNOW17 cannot directly account for the influence of many meteorological vari-

ables, such as solar and longwave radiation or wind, on snowpack processes. Surface albedo variations, in particular, have a large impact on the energy balance equations in the spring due to large solar fluxes that alter the energy balance (Jin et al. 1999a). The SAST model was also found to be more sensitive to inputs than the SNOW17 (Franz et al. 2008). Therefore, we hypothesize that the SAST, which explicitly solves the energy balance of a snowpack by incorporating multiple climate variables, will provide more accurate estimates of future uncertainties in SWE, snowmelt timing, and overall basin discharge when compared to the SNOW17, which is restricted by data insensitivity and limited inputs.

2. Methodology

a. Study sites

The Reynolds Creek Experimental Watershed (RCEW) is located in the Owyhee Mountains of southwestern Idaho (Fig. 1). The basin is characterized by a semiarid climate with precipitation ranging from approximately 230 mm in the lower elevations to over 1100 mm in the upper elevations, of which 20% and 75%, respectively, occur as snow (Hanson 2001). Data for the basin were obtained from the U.S. Department of Agriculture, Agricultural Research Service, Northwest Watershed Research Center.

Point SWE and watershed discharge forecasts were generated for a 0.39 km^2 (Pierson et al. 2001), snow-dominated subbasin with minimal relief (2024–2139 m) called Reynolds Mountain East (hereafter called the East basin). A snow pillow at the center of the East basin provided data for verification of SWE outlooks. The East basin is nested within the Tollgate basin. Discharge forecasts were generated for the Tollgate watershed, which is 54.44 km^2 in area, with an elevation range of 1398–2244 m (Pierson et al. 2001). Weir data were used to analyze watershed discharge hindcasts for the East and the Tollgate basins.

Watershed simulations were conducted in a lumped manner for both basins using the Thiessen polygon method to compute average precipitation (Franz et al. 2008). Model input data for the East basin were taken from a climate observation station located on the western edge of the basin. Model input data for the Tollgate were computed for the mean basin elevation (1837 m) using a lapse rate between two climate observation points closest to the basin (Hanson et al. 2001).

b. Models

The SNOW17 is a conceptual snow accumulation and ablation model (Anderson 1976). Energy balance equations are only explicitly used during rain on snow events when assumptions about meteorological conditions can be reasonably made. Temperature inputs are used to estimate the snowpack heat content, snow accumulation, and snowmelt. The heat deficit, liquid water retention and transmission, and the areal extent of snow cover are represented through the use of empirical equations and 10 parameters (NWS 2004). Snow is modeled as a single layer and the model requires inputs of temperature and precipitation. The model computes the SWE and snow-covered-area state variables, and outputs melt.

The SAST scheme (Jin et al. 1999a,b) is based on the physically based snow schemes of the one-dimensional snow property and process models (Jordan 1991; Anderson 1976), but has been made computationally

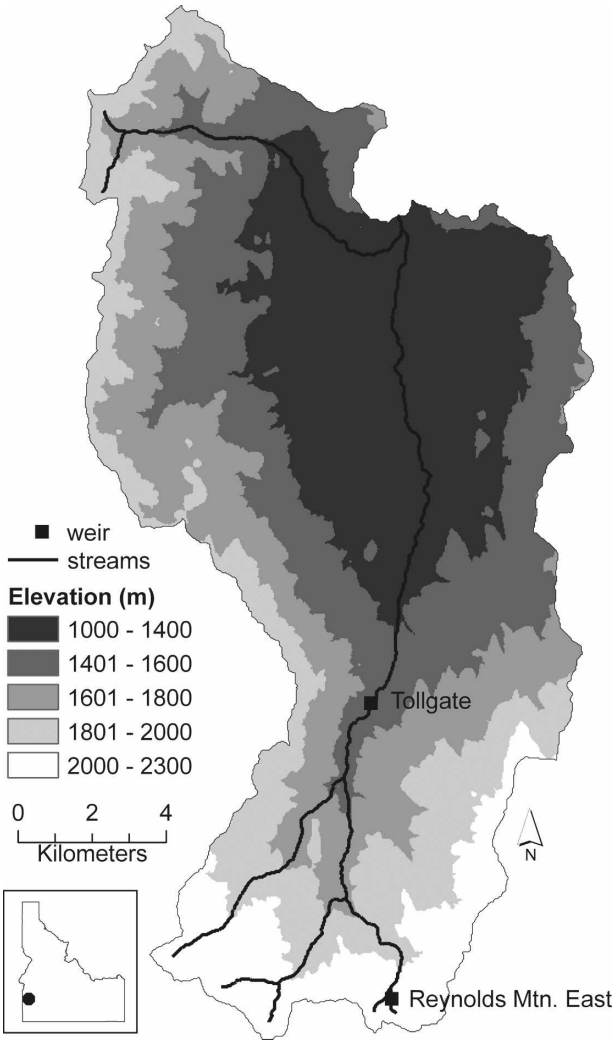

FIG. 1. Locations of the outlets of Tollgate and Reynolds Mountain East watersheds in the USDA Reynolds Creek Experimental Watershed.

simpler to allow for application to climate and hydrologic studies (Sun et al. 1999). The SAST explicitly computes snow compaction, heat conduction, grain growth, and snow melting. The model includes equations for tracking the heat content of the pack using energy balance equations, the mass balance of the snowpack, and the rate of snow compaction and snow density changes. It uses no more than three snow layers, which vary in thickness depending upon the total depth of the snow (Sun et al. 1999). SAST computes the following state and outputs variables: SWE, snow density, runoff, snow temperature profiles, and turbulent heat fluxes at the snow surface. The SAST model requires the following inputs: incoming and reflected shortwave radiation, incoming longwave radiation, air temperature, precipitation, wind speed, and relative humidity.

The SACSMA model is a saturation excess model that represents percolation, soil moisture storage, drainage, and evapotranspiration (ET) processes in a conceptual manner (Burnash et al. 1973). Inputs to the SACSMA are basin-average precipitation and/or snowmelt, and potential evaporation (PE); output is a basin-average runoff depth. NWS PE values from an adjacent forecast basin for the 16th of each month were used to linearly interpolate daily PE. The SACSMA has 16 parameters, 4 of which are typically set to default values. Runoff from the SACSMA was routed using a linear reservoir to compute basin discharge. An areal depletion curve (ADC) was used for both models to compute the snow-covered area and basin-average melt when the models were coupled to the SACSMA for watershed-scale hindcasts of discharge.

Both snow models were manually calibrated and evaluated against observations of SWE at the snow site in the East basin (Franz et al. 2008). The SACSMA was calibrated using automatic methods (Hogue et al. 2000, 2006). The snow models were run at a 1-h time step. No snow correction factor was applied and the rain–snow cutoff was set at 1°C for both models so each would receive equal amounts of snowfall.

c. *Ensemble streamflow prediction and hindcasting*

The NWS ESP system was developed as an objective method for generating long-term probabilistic streamflow outlooks (Day 1985). ESP employs operational forecast models and past meteorology to develop multiple streamflow sequences (traces) conditioned on the current basin states. A probabilistic forecast is created by statistical analysis of the multiple streamflow scenarios produced. The forecasts are based on the assumption that past meteorology is representative of possible future events (Franz et al. 2003). Forecast traces can be individually weighted based on factors such as climate; however, in the application presented here all traces are considered equally likely.

Initial basin conditions were computed by running the calibrated models throughout the entire period of record, creating a historical simulation (Franz et al. 2003), and archiving states for the first day of the desired forecasts. Hindcasts, or conditional simulations, were then generated by running the models for each historical forecast day, initializing the models from the appropriate saved state, and driving the models with input data from the historical record spanning the desired forecast period.

All available meteorological data were used for the generation of forecast traces [water years (WY) 1984–96, excluding the forecast year], resulting in 12 traces per hindcast and 13 hindcasts per forecast date. Following the schedules employed by forecasting agencies operating in the northwest United States, hindcasts were generated for 1 January (J1), 1 February (F1), 1 March (M1), 1 April (A1), and 1 May (My1) forecast dates.

A deterministic forecast predicts a single value of a variable (Croley 2000). With respect to ESP, a deterministic outlook could be obtained by choosing a single value from the ensemble such as the ensemble mean or median. A probabilistic forecast provides a predicted value(s) of a variable and the associated distribution function that reflects the likelihood of the event (Croley 2000). A probabilistic ESP forecast is produced by considering the distribution of the ensemble members. Forecast probabilities for the ESP ensembles used here were generated by distributing the ensembles into bins defined from the cumulative distribution of all available historical observations (climatology) of the prediction variable (Franz et al. 2003). For streamflow forecasts, 10 categories (10th percentile, 20th percentile, etc.) were determined from the empirical distribution of historical April–July (seasonal) discharge volumes based on 1963–96 data. (The May–July discharge volumes were used for the My1 hindcasts). The relative or cumulative frequencies of ensemble members within these categories were then computed to generate the probabilistic forecast. Ensembles of weekly mean SWEs were processed in a similar manner using observations from 1984 to 1996 to determine the SWE climatology.

d. *Forecast verification*

There are numerous forecast verification measures in the literature. We include three deterministic measures to evaluate and compare the skill of the ensemble median: the coefficient of prediction (C_p), the mean error, and the joint distribution. We also utilize two common

probabilistic measures to evaluate and compare the ensemble skill: ranked probability skill score (RPSS) and reliability.

The coefficient of prediction (C_p) (Lettenmaier 1984) is defined as

$$C_p = 1 - \frac{\mathrm{Var}(f_{\mathrm{med}} - o)}{\mathrm{Var}(o)}, \quad (1)$$

where o is the observation, f_{med} is the forecast (ensemble median), and $\mathrm{Var}(\cdot)$ is the variance of each of the respective variables. A value of C_p equal to 1.0 indicates a perfect forecast and a C_p less than zero indicates that the mean of the observations is a better predictor than the forecast. Mean error (ME) is the average difference between the forecasts and the observations:

$$\mathrm{ME} = \frac{1}{n}\sum_{k=1}^{n} f_{\mathrm{med},k} - o_k, \quad (2)$$

where n is the number of forecasts.

The joint distribution of the forecast and the observation [$p(f_{\mathrm{med}}, o)$] can be analyzed through the use of a scatterplot, where perfect forecasts fall along the 1:1 line. The joint distribution provides information about the forecasts, the observations, and the relationship between the forecasts and observations (Murphy and Winkler 1987). Although common practice, choosing one value from an ensemble forecast assumes that the probability of the single value is 1. The skill of the likelihood estimates and information about ensemble spread cannot be accurately assessed by evaluating a deterministic value. Therefore, we plot the 10th and 90th percentiles from the ensembles on the joint distribution diagrams and include the following statistics.

The ranked probability score (RPS_L; Epstein 1969; Wilks 1995; Müller et al. 2005) is a measure of the overall accuracy of multicategory forecasts. RPS_L is defined as the mean square error of the cumulative probability distribution of the forecasts (F_m) and the observations (O_m):

$$\mathrm{RPS}_L = \sum_{m=1}^{J} |F_m - O_m|^L, \quad (3)$$

$$F_m = \sum_{j=1}^{m} f_j,\ m = 1, \ldots J, \quad \text{and} \quad (4)$$

$$O_m = \sum_{j=1}^{m} o_j,\ m = 1, \ldots J, \quad (5)$$

where f_j is the relative frequency of the forecast traces, o_j is the relative frequency of the observations, J is the number of event categories, and L is the norm. In the standard definition of the RPS_L (Wilks 1995), $L = 2$. The observation occurs in only one of the categories, which is given a value of 1; the remaining categories are given a value of 0 (Wilks 1995). For a group of n forecasts, the RPS_L is the average ($\overline{\mathrm{RPS}_L}$) of the n RPS_L scores:

$$\overline{\mathrm{RPS}_L} = \frac{1}{n}\sum_{k=1}^{n} \mathrm{RPS}_{L,k}. \quad (6)$$

An RPS_L score of 0 indicates a perfect forecast; nonperfect forecasts have positive RPS_L values. The RPS_L skill score (RPSS_L) is used to evaluate the relative skill of a set of forecasts ($\overline{\mathrm{RPS}}_{L,f}$) to the skill of a reference forecast (climatology is used here)($\overline{\mathrm{RPS}}_{L,r}$):

$$\mathrm{RPSS}_L = \left(1 - \frac{\overline{\mathrm{RPS}}_{L,f}}{\overline{\mathrm{RPS}}_{L,r}}\right) \times 100\%. \quad (7)$$

A positive (negative) RPSS_L indicates that the forecasts performed better (worse) on average than the reference forecasts. $\mathrm{RPSS}_{L=2}$ has been shown to be negatively biased for small ensemble sizes (Kumar et al. 2001; Müller et al. 2005). Here, $L = 1$ is used to reduce bias problems associated with small ensembles, following suggestions by Müller et al. (2005). $\mathrm{RPSS}_{L=1}$ and $\mathrm{RPSS}_{L=2}$ were compared for the hindcasts studied, and $RPSS_{L=1}$ scores were higher with no instances of negative values as with $\mathrm{RPSS}_{L=2}$. The relative skill levels between the two modeling systems were similar for both forms of RPSS_L.

Reliability is the conditional distribution of the observation (o) given the forecast (f) $p(o|f)$ (Murphy and Winkler 1992). Reliability diagrams plot the observed relative frequency as a function of forecast probability (Wilks 1995); a perfectly reliable set of forecasts will plot along a 1:1 line. Forecasts that plot to the left of the 1:1 line are underpredicting the observations; forecasts that plot to the right of the 1:1 line are overpredicting the observations.

Relative frequency diagrams of the forecast probability $p(f_j)$ are provided as an inset in the reliability plots to indicate the sharpness of the forecasts (or resolution). As forecasts become sharper, the probability is more frequently assigned to the extreme probability categories (i.e., 0%–20% or 80%–100%) (Murphy and Winkler 1987). The frequency diagrams can be used to indicate which forecast probability categories may be susceptible to being skewed by outliers due to small sample size.

3. Results and discussion

Streamflow hindcast results for the SNOW17 coupled to the SACSMA are referred to as SNOW–

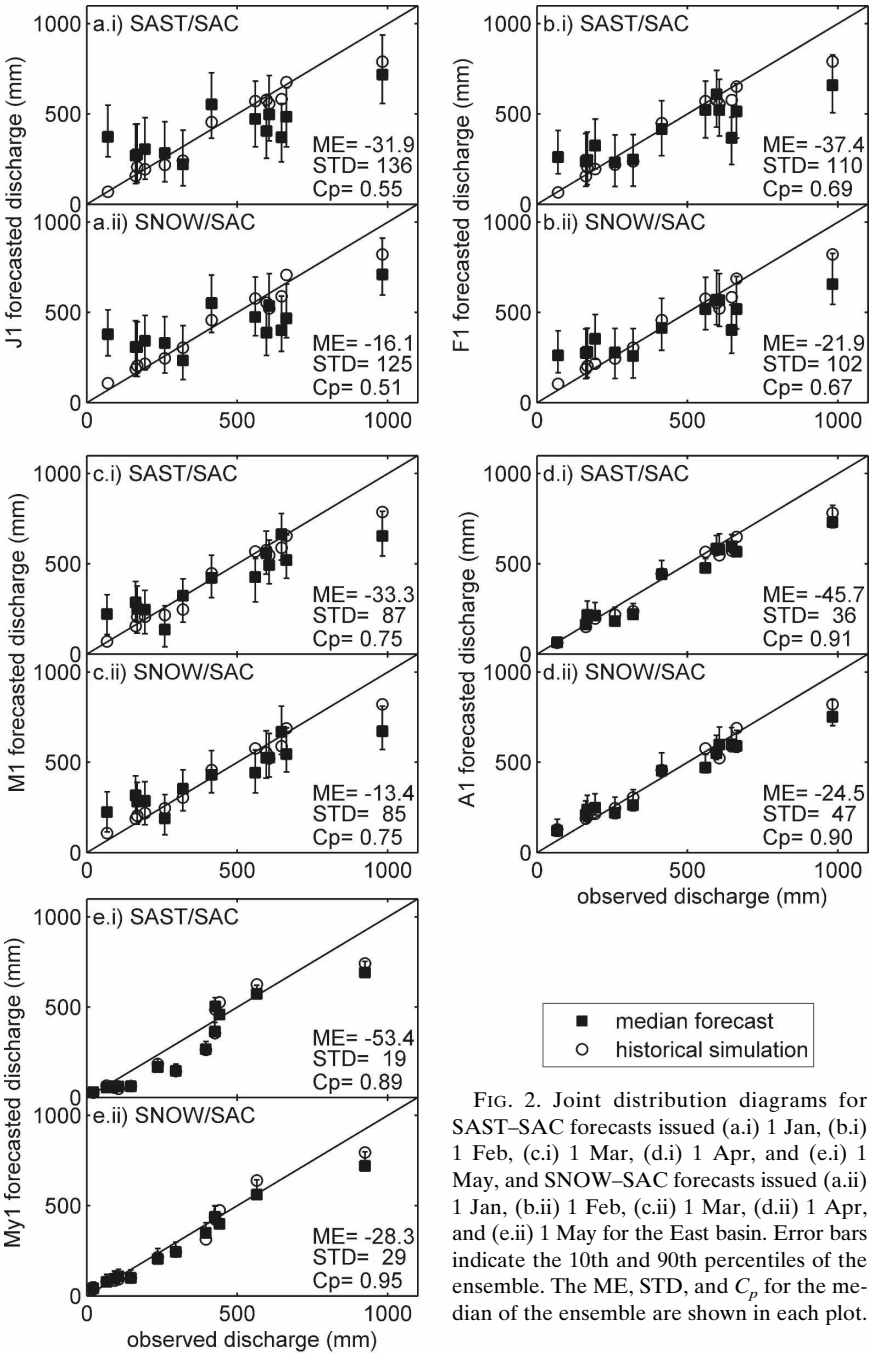

FIG. 2. Joint distribution diagrams for SAST–SAC forecasts issued (a.i) 1 Jan, (b.i) 1 Feb, (c.i) 1 Mar, (d.i) 1 Apr, and (e.i) 1 May, and SNOW–SAC forecasts issued (a.ii) 1 Jan, (b.ii) 1 Feb, (c.ii) 1 Mar, (d.ii) 1 Apr, and (e.ii) 1 May for the East basin. Error bars indicate the 10th and 90th percentiles of the ensemble. The ME, STD, and C_p for the median of the ensemble are shown in each plot.

SAC, and results for the SAST coupled to the SACSMA are referred to as SAST–SAC. Seasonal discharge forecasts are analyzed for total basin outflow occurring from 1 April through 31 July for each forecast date, with the exception of the My1 forecast for which the total outflow occurring from 1 May through 31 July is analyzed. The seasonal discharge forecasts are evaluated using C_p, ME, joint distribution, $\text{RPSS}_{L=1}$, and reliability. Weekly SWE ensembles are evaluated using C_p, joint distribution, and $\text{RPSS}_{L=1}$.

a. Discharge volume forecasts

The joint distribution diagrams for forecasts of the East basin discharge are shown in Fig. 2. For both models, the correlation between the median forecasts and the observations improves (falling along the 1:1 line)

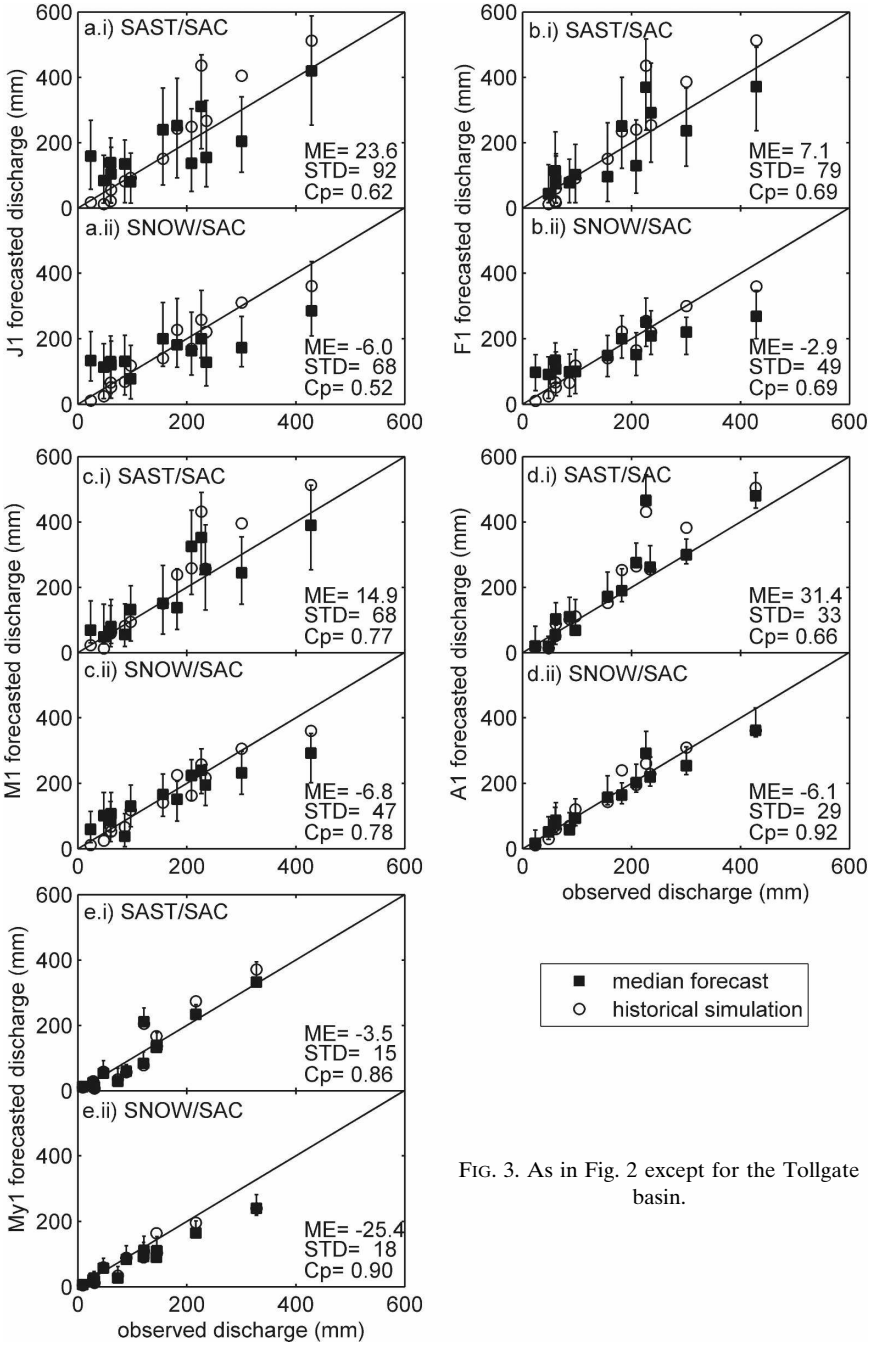

FIG. 3. As in Fig. 2 except for the Tollgate basin.

and the ensemble spread decreases (reduced error bars) from the J1 forecast (Fig. 2a) to the My1 forecast (Fig. 2e). The C_p also improves with decreasing lead time as indicated by the increasing values. The SNOW–SAC forecasts in the East basin for My1 are near perfect with a C_p of 0.95. The C_p of the ensemble median are slightly higher in the Tollgate basin for the J1, F1, and M1 forecasts (shown in Figs. 3a–c, respectively). The C_p values improve with decreasing lead time, with the exception of the A1 (Fig. 3d) forecast for the SAST–SAC.

The joint distributions diagram reveals one potential outlier (1989) in the SAST–SAC A1 forecasts in the Tollgate basin. The 1989 median forecast overestimates the seasonal discharge by 50%. With the A1 1989 forecast removed from the SAST–SAC set, the C_p value increases to 0.94, slightly higher than the SNOW–SAC score. The SAC–SAC simulation for 1989 had very little snowmelt until mid-April, resulting in an oversimulation of spring streamflow. Because the hindcasts were initialized with modeled SWE, the SWE state in

the model was erroneously high on 1 April and streamflow was overpredicted. On average, the median forecasts from both models were good predictors of streamflow for the A1 and My1 forecasts in both basins, with the SNOW17 model scoring slightly higher.

The forecast ensemble spread is quite similar for both models in the East basin as indicated by the joint distribution diagram and the ensemble standard deviations (STDs; Fig. 2). In the Tollgate basin, the SAST–SAC ensembles have a significantly higher standard deviation and a noticeably larger spread compared to the SNOW–SAC ensembles for the J1, F1, and M1 forecasts (Fig. 3). There are three SNOW–SAC forecasts where the observations are not captured within the 80% probability range for the J1 (Fig. 3a) and F1 (Fig. 3b) forecast dates. In contrast, the observations are within the 10% and 90% bounds of the SAST–SAC ensembles for all but one forecast for the J1 and F1 forecast dates. The additional climate forcings required by the SAST appear to provide additional and more representative information about future snowmelt (and subsequently streamflow) probability in the larger Tollgate basin for early season forecasts.

The F1 and M1, and to a lesser extent the J1, forecasts in the Tollgate basin hint at the problem with analyzing only the median trace from the ensemble. In both cases, the forecasts from each model have the same or nearly the same C_p value and the mean errors are arguably very close. However, the SAST–SAC has clearly more spread in the ensemble and captures the observation within the 80% probability range shown more often than the SNOW–SAC, indicating higher accuracy. These differences in forecast characteristics are not reflected in the deterministic scores. As an additional example, the SAST–SAC M1 (Fig. 2c.i) forecast would be more accurate than the My1 (Fig. 2e.i) forecasts in the East basin because the M1 ensembles capture the observation more often than the My1 ensembles, which have a very low spread. However, the C_p for the My1 forecast is higher, which suggests more skill. There appears to be little correlation between ME and C_p, making it difficult to interpret the relationship between the two metrics. The higher ensemble accuracy in the M1 SAST forecast compared to the My1 SAST forecast is captured by the $\mathrm{RPSS}_{L=1}$ (Fig. 4a).

The sensitivity of the RPSS_{L1} value was examined assuming a conservative estimate of uncertainty (25% normally distributed noise) in the observed streamflow. A significant difference in RPSS_{L1} skill was assumed if the 5%–95% confidence intervals did not overlap. Because there was considerable overlap in the confidence intervals, the probabilistic hindcasts from both models are considered to be equally skillful in both basins with

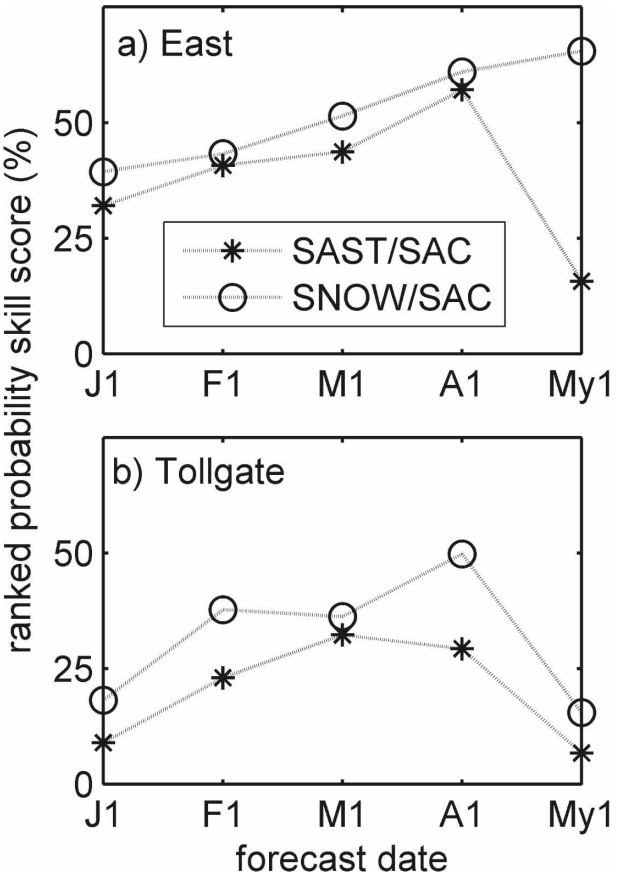

FIG. 4. $\mathrm{RPSS}_{L=1}$ for the 1 Jan–1 May seasonal discharge volume hindcasts for the (a) East and (b) Tollgate basins.

the exception of the My1 forecast in the East basin (Fig. 4a), and the F1 and A1 forecasts in the Tollgate basin (Fig. 4b). The $\mathrm{RPSS}_{L=1}$ does not adequately reveal the fact that the SAST ensemble more often contains the observation compared to the SNOW17 (most prominent in the F1 and M1 forecasts in the Tollgate basin). Because the $\mathrm{RPSS}_{L=1}$ is sensitive to distance, the SNOW17 model benefits, on average, from having smaller ensemble spread in those years where the probability estimates are accurate, which produces higher $\mathrm{RPSS}_{L=1}$ scores.

Hindcasts from both models were more skillful than climatology for the forecast dates (Fig. 4). The behavior of the SAST–SAC hindcasts was similar in both basins, with improved skill until A1, and then decreased skill in the latter part of the forecast season. The SNOW–SAC had increasing $\mathrm{RPSS}_{L=1}$ values throughout the season in the East basin. The SNOW–SAC My1 ensembles in the Tollgate basin tend to underpredict the observations (Fig. 3), resulting in a lower $\mathrm{RPSS}_{L=1}$ for this period.

The general trend toward improved $\mathrm{RPSS}_{L=1}$ later in

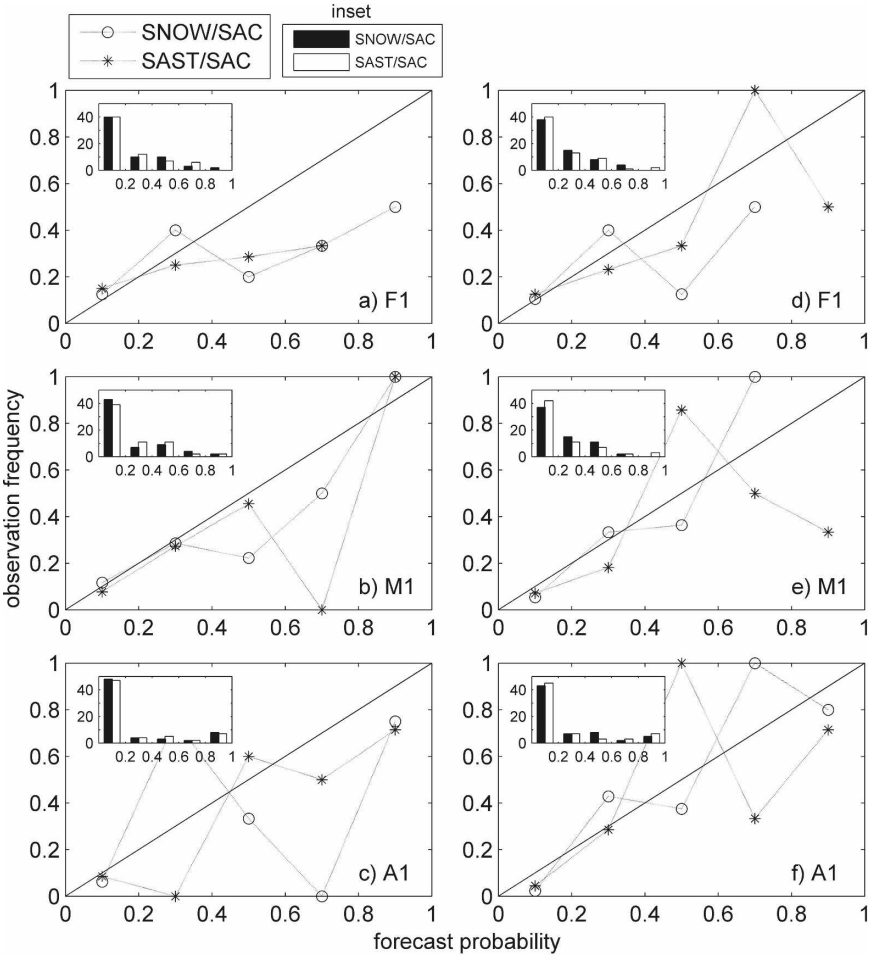

FIG. 5. Nonexceedance seasonal discharge volume hindcasts for the (a) 1 Feb, (b) 1 Mar, and (c) 1 Apr forecast dates in the East basin and (d)–(f) for the same dates in the Tollgate basin. The forecast frequency diagrams are shown in the inset, where the *y* axis is the forecast frequency and the *x* axis is the forecast probability.

the forecast season (up until ~A1) as lead time decreased is a trend that was observed in other forecast studies in the western United States (Franz et al. 2003; Pagano et al. 2004). Franz et al. (2003) also observed a decline in skill after major melting in several watersheds studied. In most years in the study basins, a significant amount of melting took place prior to 1 May. The climate inputs produce a larger amount of variability in the predicted hydrographs when the snowpack is at or near peak accumulation (Figs. 2 and 3). Late in the spring, the influence of the snowpack is largely removed and the meteorological variability has a comparatively greater influence on the prediction. With insufficient variability in the historical meteorology, the spread of the ensemble decreases and the chance of capturing the observation decreases. In the operational forecast environment, current watershed observations are likely to be accounted for as part of the forecast process, thus improving the likelihood that the forecasts will be accurate even during major melting, as was the case in the Pagano et al. (2004) study. The influence of initializing models with observed SWE will be explored in the next section.

Select reliability diagrams are presented in Fig. 5. The forecast resolution (the distribution of forecast probability) is similar between the two models and <40% nonexceedance probabilities occurred, most often indicating that the ensemble members were fairly well-distributed across the forecast categories (Fig. 5, insets). At these low probabilities, the forecasts are generally reliable (i.e., results plot along the one-to-one line), with the exception of the East basin A1 forecasts (Fig. 5c). Forecasted events with >60% probability tend to be overpredicted, particularly for the East basin. Forecasts issuing 80%–100% probability in the East basin (Figs. 5b and 5c) and the Tollgate basin (Fig.

5f) are fairly reliable despite the low forecast counts in these bins, which can lead to a negative influence by outliers.

The root-mean-squared errors (RMSEs) of the calibrated SNOW–SAC and SAST–SAC were 0.13 (0.08) and 0.15 (0.10) mm day^{-1}, respectively, in the East basin (Tollgate basin). The SAST–SAC tended to have poorer RMSEs and other model simulation statistics, and the minor differences in model accuracy likely contributed to the poorer $RPSS_{L=1}$ forecast scores. However, the C_p and reliability results are very similar, and it appears that calibration plays a secondary role in forecast skill, with initial conditions and meteorological input being more important. Future investigations regarding model state updating of the SACSMA state should reveal more information regarding the relative importance of calibration, initial conditions, and climate forcing.

b. SWE forecasts

The primary purpose of the current study is to determine how the two snow models perform relative to one another for streamflow prediction. Therefore, it is important to analyze how the models simulate and predict the SWE state. The predicted weekly mean SWE forecasts from the East basin are analyzed using C_p and $RPSS_{L=1}$.

Ensemble predictions of weekly mean SWE from both models (not shown) had $RPSS_{L=1}$ scores between 50% and 80% at the start of the forecast and continually decrease to some point within the forecast window (approximately the last quarter of the forecast window) when the skill becomes worse than climatology (below 0% $RPSS_{L=1}$). The differences in the $RPSS_{L=1}$ values are greatest at the start of the forecast window, with the SNOW17 $RPSS_{L=1}$ values ranging between 7% (J1) and 15% (F1 and A1) higher than those of the SAST. However, these ranges are likely within the verification uncertainty. The SNOW17 J1, F1, and M1 $RPSS_{L=1}$ values are higher than those of the SAST throughout much of the forecast window. The A1 and My1 forecasts and forecast skills tend to be nearly equal between the two models after weeks 4 and 2, respectively.

The mean daily SWE errors of the calibrated SNOW17 and SAST models were 26.7 and 34.2 mm, respectively. The hindcasts presented thus far were impacted in part by the accuracy of the model calibration because, as stated previously, initial states were saved during the historical simulations and were not updated for the hindcasting. The SAST simulation tended to have a late onset of melt resulting, in overestimated spring SWE values (Franz et al. 2008). After the initial onset of melt, melting occurred rapidly. Therefore, average errors in the SAST initial SWE states were positive for J1 to M1 forecast dates and negative for A1 and My1 forecast dates when compared to the snow pillow observations in the East basin (Fig. 6). On average, the initial SWE states in the SAST model historical run (which generates initial conditions for the ensemble) had larger errors than did the SNOW17. The decreasing relative SWE errors for the forecast dates from F1 to M1 coincide with increasing $RPSS_{L=1}$ values. The shift to underestimated initial conditions in the SAST and the SNOW17 models during the spring coincide with decreased $RPSS_{L=1}$ values for seasonal discharge in the East basin. Overestimated (underestimate) initial conditions correspond with overestimated (underestimated) seasonal discharge (Fig. 2).

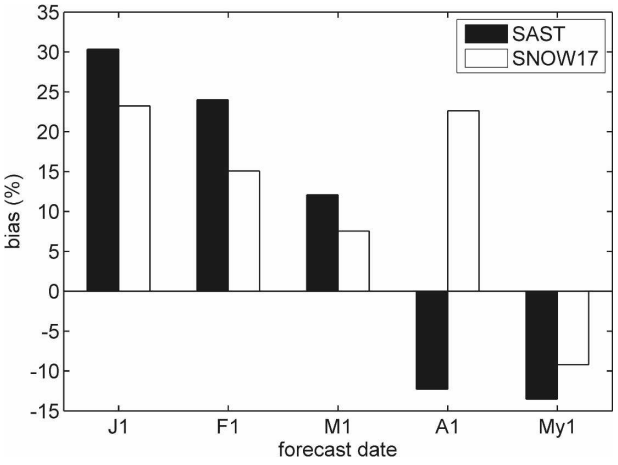

FIG. 6. Average percent bias in the initial conditions of SWE (first day of the forecasts) in the SAST and SNOW17 models.

To investigate the impact of the SWE errors on the forecasted SWE ensembles, both snow models were reinitialized with the observed SWE from the East basin snow pillow data and the SWE hindcasting was repeated. Updating was done using direct substitution. In the case of the SNOW17 model, the water equivalent (WE) state was updated by subtracting the liquid water content (LIQW) state from the observed SWE (together these equal the total SWE). For the SAST model, each of the three snow layers were adjusted, keeping their relative water contents the same; the layer thicknesses were not changed. All other snow model states were left unchanged.

This simple updating procedure improved the $RPSS_{L=1}$ of both the SNOW17 and SAST SWE forecasts for 1–4 weeks into the forecast window (Fig. 7). After updating, the $RPSS_{L=1}$ scores from the two models were equal at the start of the forecast windows and ranged between 83% (My1; see Fig. 7e) and 88% (F1; see Fig. 7b). After 4 weeks for J1 (Fig. 7a), F1 (Fig. 7b),

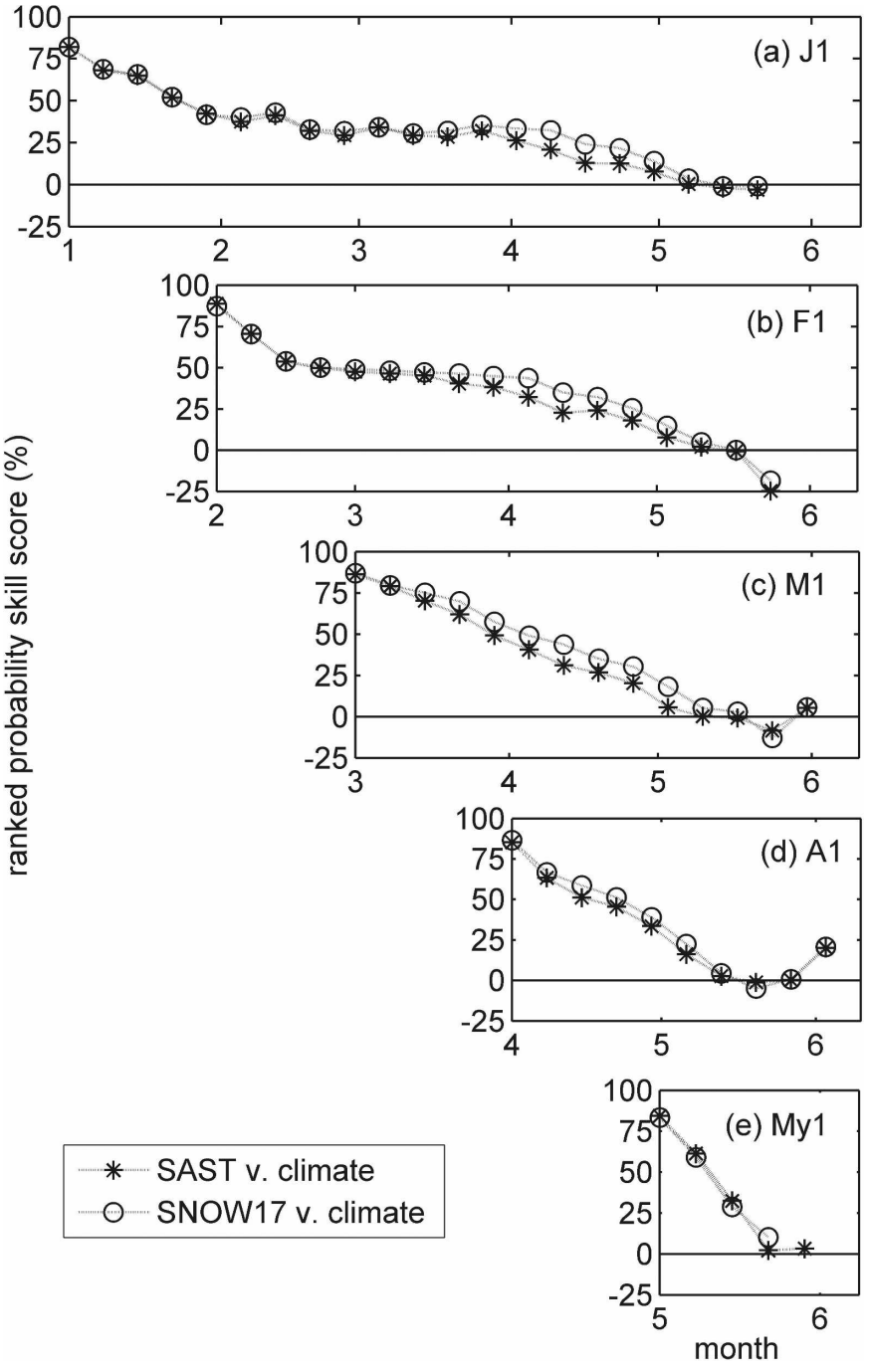

FIG. 7. $RPSS_{L=1}$ for hindcasts of weekly SWE at the East basin snow pillow observation site for forecast dates of (a) 1 Jan, (b) 1 Feb, (c) 1 Mar, (d) 1 Apr, and (e) 1 May, generated after the initial conditions of the SWE in the models were updated to the observed values.

and M1 (Fig. 7c) forecasts, and 2 weeks for A1 (Fig. 7d) and My1 (Fig. 7e) forecasts, the skill of the predictions return to the preupdated values. This indicates that the initial conditions are dominating the ensemble skill in the early part of the forecast window relative to the climate data; however, climate data become increasingly important after several weeks into the forecast, particularly during the melt period.

During the early part of the forecast season, the average number of days between the earliest and latest

TABLE 1. Standard deviation of ensemble members (spread) from the updated SNOW17 and SAST SWE hindcasts for predicted maximum SWE (mm) and the number of days to the predicted melt of the snowpack at the East basin snow pillow observation site.

Forecast date	Predicted max SWE (mm)		Time to predicted melt (days)	
	SNOW17	SAST	SNOW17	SAST
1 Jan	12.52	13.13	2.29	2.10
1 Feb	11.16	10.71	2.07	1.80
1 Mar	5.83	5.92	1.55	1.79
1 Apr	1.62	1.86	1.60	4.01
1 May	1.04	1.87	3.15	3.00

predicted complete melt of the pack in the ensembles is around 60 days for the SAST and 50 days for the SNOW17 (not shown). For the My1 forecast this range dropped to 20 days for both models. The effect of updating the SWE state on the range of days between the predicted complete melt was examined and found to have no significant impact. Therefore, the uncertainty in the timing of the snowmelt is highly dependent upon the input and not the initial SWE state; this is supported by a rapid decrease in $\mathrm{RPSS}_{L=1}$ by week 2 during the melt period.

The updated SWE ensemble spread for the maximum predicted SWE and the number of days to the predicted melt are provided in Table 1. The spread of the predicted maximum SWE by the SAST is greater than the SNOW17 for four of the five forecast dates. Although there was some indication that the additional climate inputs improved the spread of the Tollgate basin discharge forecast, it is unclear that the climate variables used by the SAST had an effect on the SWE ensemble spread.

The C_p (Fig. 8) is the average of all weekly C_p values computed from the median weekly SWE trace for each forecast date. The C_p values were highest for the J1, F1, and M1 forecast dates; and the SNOW17 had a slightly higher score than the SAST. The C_p scores decrease considerably in the A1 and My1 forecasts for both models. The winter C_p scores are lower after updating the initial conditions in the models; however, the A1 and My1 scores are improved (Fig. 8). Most SWE corrections that were made decreased the initial SWE value. The joint distribution of the updated weekly SWE median trace and observed SWE (not shown) showed a very slight increased tendency of the updated median forecasts of both models to underpredict weekly SWE for the J1, F1, and M1 SWE ensembles, explaining the lower C_p score. The correlation between the My1 median forecast and the observations was better after updating.

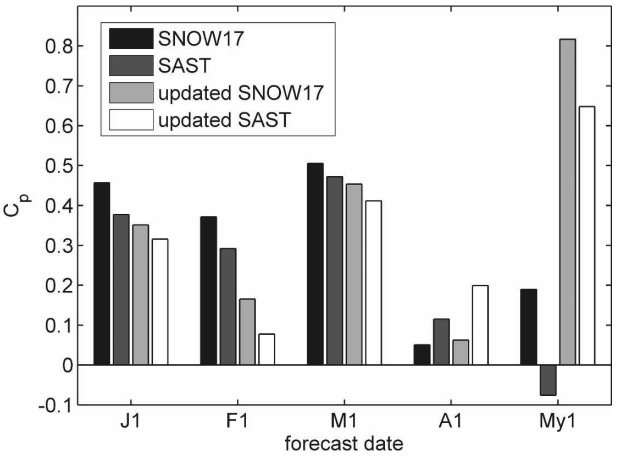

FIG. 8. The C_p for the median forecast of the mean weekly SWE at the East basin snow pillow observation site.

SACSMA states cannot be updated by simple substitution of observed variables because model states are not directly related to physical observations. Currently, no systematic state updating method is applied operationally. In addition, there is minimal documentation about the state modifications that are performed, so evaluation of historical adjustments is not possible and adjustments are not repeatable. Given the dominance of the initial conditions on the ensemble predictions and the skill of the 12-member ensemble (Fig. 7), ESP combined with objective state updating may be appropriate for regions with limited climate records.

Basin-average precipitation was computed using the Thiessen polygon method with the two precipitation gauges in the East basin and nine precipitation gauges in the Tollgate basin. Given the minimal relief in the East basin, the use of the Thiessen polygon method is reasonable. Although the Tollgate basin has more substantial relief, precipitation gauges are well distributed, including five gauges in the highest elevations, providing good representation of precipitation throughout the watershed. Precipitation errors, as well as basin-average estimates of other climate variables and SWE observations, impact the forecast scores. However, both models were applied such that they received an equal mass of precipitation; therefore, errors associated with uncertain precipitation input would be reflected equally in verification scores from each model.

4. Conclusions

More rigorous model validation is needed within the hydrologic community, especially within the context of operational transferability. In the hindcasting method applied here, the initial model conditions were first determined from historical model runs. Since the models

and data are not perfect, there is uncertainty in the estimated initial conditions for the forecast. Accurate initial conditions were shown to be more important for ESP forecast skill than the type of snow model used. Without updating, SWE and discharge predictions from the SAST model were shown to be slightly less skillful than the SNOW17. After updating, the SAST SWE hindcasts were equally skillful as the SNOW17 based on $RPSS_{L=1}$. Regardless of the type of snow model used, objective state updating techniques will be a necessary part of an advanced forecast system, and arguably will have a greater impact on forecast skill than the chosen snow model. Data assimilation has been shown to improve hydrologic simulations in the SACSMA (Seo et al. 2003; Vrugt et al. 2006). Future studies will focus on data assimilation method development for both the SNOW17 and the SACSMA and investigate the impact of state updating (of multiple model states) on hindcast skill.

Climate appears to dominate the latter part of the forecast, as the effects of the initial condition update diminish after several weeks, depending upon the time of year. In the springtime, meteorological inputs to the models become increasingly important because snowmelt is highly variable during this time period. Therefore, updating model states improved the $RPSS_{L=1}$ in only the first 1–2 weeks for the M1, A1, and My1 forecasts compared to 4 weeks seen in the earlier season forecasts. Forecast skill in the latter part of the window remained unchanged, and is primarily a function of the climate input. Additional climate variables required by the SAST appear to improve the spread and accuracy of forecasts in the larger basin (Tollgate). The degree of spread of the ensemble forecasts from both models were very similar in the East basin. The climate time series in this basin may have lacked enough variability to produce any noticeable difference between the hindcasts from the two snow models.

Advanced forecast model evaluation must go beyond the traditional simulation experiments to increase its relevancy to forecasting applications. A relatively simple hindcasting application that requires only the data and tools that are used in a model evaluation study has been shown as an additional, but critical, step to a traditional model comparison study. Although limited by the available data, this study has met the criteria put forth by Kirchner et al. (1996). The SAST has been evaluated for simulation of the desired variables SWE and discharge; has been compared to the SNOW17, an appropriate benchmark; and has been applied with the SACSMA for generating ESP hindcasts, which is consistent with current forecasting procedures. The results indicate that the SAST is skillful for seasonal predictions, that this skill is comparable with the SNOW17's skill, and that ESP forecasts from both models can be improved if SWE state updating is conducted. A multimodeling method, which would allow the various skills of both snow models to be exploited, is a logical approach and will also be investigated.

Acknowledgments. Primary financial support provided by the NASA GEWEX program (Grant NNG04GM24G), and NASA EOS (Grant NAG5-11044) is appreciated. Financial support provided by the University of California, Irvine's Newkirk Center for Science and Society is also greatly appreciated. Data for the Reynolds Creek Watershed were provided by the U.S. Department of Agriculture's Northwest Watershed Research Center. We would also like to thank the University of California, Irvine, for support of this work.

REFERENCES

Anderson, E. A., 1973: National Weather Service River Forecast System—Snow accumulation and ablation model. NOAA Tech. Memo. NWS Hydro-17, 217 pp.

——, 1976: A point energy and mass balance model of a snow cover. NOAA Tech. Rep. NWS 19, 150 pp.

Bradley, A. A., S. S. Schwartz, and T. Hashino, 2004: Distributions-oriented verification of ensemble streamflow predictions. *J. Hydrometeor.,* **5,** 532–545.

Burnash, R. J., R. L. Ferral, and R. A. McGuire, 1973: A generalized streamflow simulation system conceptual: Modeling for digital computers. Joint Federal–State River Forecast Center, Sacramento, CA, 204 pp.

Carpenter, T. M., and K. P. Georgakakos, 2001: Assessment of Folsom Lake response to historical and potential future climate scenarios: 1. Forecasting. *J. Hydrol.,* **249,** 148–175.

Clark, M. P., and L. E. Hay, 2004: Use of medium-range numerical weather prediction model output to produce forecasts of streamflow. *J. Hydrometeor.,* **5,** 15–32.

Croley, T. E., 2000: *Using Meteorology Probability Forecasts in Operational Hydrology.* American Society of Civil Engineers (ASCE) Press, 206 pp.

Day, G. N., 1985: Extended streamflow forecasting using NWSRFS. *J. Water Resour. Plann. Manage.,* **111,** 157–170.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.,* **8,** 985–987.

Franz, K. J., H. C. Hartmann, S. Sorooshian, and R. Bales, 2003: Verification of National Weather Service ensemble streamflow predictions for water supply forecasting in the Colorado River basin. *J. Hydrometeor.,* **4,** 1105–1118.

——, N. Ajami, J. Schaake, and R. Buizza, 2005: Hydrologic ensemble prediction experiment focuses on reliable forecasts. *Eos, Trans. Amer. Geophys. Union,* **86,** 239.

——, T. S. Hogue, and S. Sorooshian, 2008: Operational snow modeling: Addressing the challenges of an energy balance model for National Weather Service forecasts. *J. Hydrol.,* **360,** 48–66.

Gomez-Landesa, E., and A. Rango, 2002: Operational snowmelt runoff forecasting in the Spanish Pyrenees using the snowmelt runoff model. *Hydrol. Processes,* **16,** 1583–1591.

Hamlet, A. F., and D. P. Lettenmaier, 1999: Columbia River streamflow forecasting based on ENSO and PDO climate signals. *J. Water Resour. Plann. Manage.,* **125,** 333–341.

Hanson, C. L., 2001: Long-term precipitation database, Reynolds Creek Experimental Watershed, Idaho, United States. *Water Resour. Res.,* **37,** 2831–2834.

——, D. Marks, and S. S. van Vactor, 2001: Long-term climate database, Reynolds Creek Experimental Watershed, Idaho, United States. *Water Resour. Res.,* **37,** 2839–2841.

Hogue, T. S., S. Sorooshian, H. Gupta, A. Holz, and D. Braatz, 2000: A multistep automatic calibration scheme for river forecasting models. *J. Hydrometeor.,* **1,** 524–542.

——, H. Gupta, and S. Sorooshian, 2006: A "user-friendly" approach to parameter estimation in hydrologic models. *J. Hydrol.,* **320,** 202–217.

Jin, J., X. Gao, S. Sorooshian, Z.-L. Yang, R. Bales, R. E. Dickinson, S. F. Sun, and G. X. Wu, 1999a: One-dimensional snow water and energy balance mode for vegetated surfaces. *Hydrol. Processes,* **13,** 2467–2482.

——, ——, Z.-L. Yang, R. C. Bales, S. Sorooshian, and R. E. Dickinson, 1999b: Comparative analyses of physically based snowmelt models for climate simulations. *J. Climate,* **12,** 2643–2657.

Jordan, R., 1991: A one-dimensional temperature model for a snow cover. Cold Regions Research and Engineering Laboratory Special Rep. 91-16, U.S. Army Corps of Engineers, 49 pp.

Kirchner, J. W., R. P. Hooper, C. Kendall, C. Neal, and G. Leavesley, 1996: Testing and validating environmental models. *Sci. Total Environ.,* **183,** 33–47.

Kumar, A., A. G. Barnston, and M. P. Hoerling, 2001: Seasonal predictions, probabilistic verifications, and ensemble size. *J. Climate,* **14,** 1671–1676.

Lettenmaier, D. P., 1984: Limitations on seasonal snowmelt forecast accuracy. *J. Water Resour. Plann. Manage.,* **110,** 255–269.

McEnery, J., J. Ingram, Q. Y. Duan, T. Adams, and L. Anderson, 2005: NOAA's Advanced Hydrologic Prediction Service—Building pathways for better science in water forecasting. *Bull. Amer. Meteor. Soc.,* **86,** 375–385.

Müller, W. A., C. Appenzeller, F. J. Doblas-Reyes, and M. A. Liniger, 2005: A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small sample sizes. *J. Climate,* **18,** 1513–1523.

Murphy, A. H., and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.,* **115,** 1330–1338.

——, and ——, 1992: Diagnostic verification of probability forecasts. *Hydrol. Processes,* **7,** 435–455.

NWS, cited 2004: The National Weather Service River Forecast System user's manual. [Available online at http://www.nws.noaa.gov/ohd/hrl.]

Pagano, T., D. Garen, and S. Sorooshian, 2004: Evaluation of official western U.S. seasonal water supply outlooks, 1922–2002. *J. Hydrometeor.,* **5,** 896–909.

Perkins, T. R., 1988: Seasonal streamflow forecasting in the upper Rio Grande basin by incorporating the use of SNOTEL data in the SSARR hydrologic model. *Proc. 56th Annual Western Snow Conf.,* Kalispell, MT, Western Snow Conference, 58–69.

Pierson, F. B., C. W. Slaughter, and Z. K. Cram, 2001: Long-term stream discharge and suspended-sediment database, Reynolds Creek Experimental Watershed, Idaho, United States. *Water Resour. Res.,* **37,** 2857–2861.

Rango, A., 1989: Evolution of a research-oriented snowmelt-runoff simulation model into an operational forecasting tool. *Proc. 57th Annual Western Snow Conf.,* Fort Collins, CO, Western Snow Conference, 45–51.

——, and J. Martinec, 1994: Model accuracy in snowmelt-runoff forecasts extending from 1 to 20 days. *Water Resour. Bull.,* **30,** 463–470.

Roulin, E., and S. Vannitsem, 2005: Skill of medium-range hydrological ensemble predictions. *J. Hydrometeor.,* **6,** 729–744.

Schaake, J., K. Franz, A. Bradley, and R. Buizza, 2006: The Hydrologic Ensemble Prediction Experiment (HEPEX). *Hydrol. Earth Syst. Sci. Discuss.,* **3,** 3321–3332.

Seo, D. J., V. Koren, and N. Cajina, 2003: Real-time variational assimilation of hydrologic and hydrometeorological data into operational hydrologic forecasting. *J. Hydrometeor.,* **4,** 627–641.

Slaughter, C. W., D. Marks, G. N. Flerchinger, S. S. Van Vactor, and M. Burgess, 2001: Thirty-five years of research data collection at the Reynolds Creek Experimental Watershed, Idaho, United States. *Water Resour. Res.,* **37,** 2819–2823.

Sun, S. F., J. Jin, and Y. K. Xue, 1999: A simple snow–atmosphere–soil transfer model. *J. Geophys. Res.,* **104** (D16), 19 587–19 597.

Vrugt, J. A., H. V. Gupta, and B. O. Nuallain, 2006: Real-time data assimilation for operational ensemble streamflow forecasting. *J. Hydrometeor.,* **7,** 548–565.

Welles, E., S. Sorooshian, G. Carter, and B. Olsen, 2007: Hydrologic verification—A call for action and collaboration. *Bull. Amer. Meteor. Soc.,* **88,** 503–511.

Werner, K., D. Brandon, M. Clark, and S. Gangopadhyay, 2004: Climate index weighting schemes for NWS ESP-based seasonal volume forecasts. *J. Hydrometeor.,* **5,** 1076–1090.

——, ——, ——, and ——, 2005: Incorporating medium-range numerical weather model output into the ensemble streamflow prediction system of the National Weather Service. *J. Hydrometeor.,* **6,** 101–114.

Westrick, K. J., P. Storck, and C. F. Mass, 2002: Description and evaluation of a hydrometeorological forecast system for mountainous watersheds. *Wea. Forecasting,* **17,** 250–262.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.

Wood, A. W., E. P. Maurer, A. Kumar, and D. P. Lettenmaier, 2002: Long-range experimental hydrologic forecasting for the eastern United States. *J. Geophys. Res.,* **107,** 4429, doi:10.1029/2001JD000659.

——, A. Kumar, and D. P. Lettenmaier, 2005: A retrospective assessment of National Centers for Environmental Prediction climate model-based ensemble hydrologic forecasting in the western United States. *J. Geophys. Res.,* **110,** D04105, doi:10.1029/2004JD004508.